Requirements of a Remote Repository

Kevin Bradley

Abstract

In 2007 UNESCO Memory of the World published a report: "Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development". The paper argued for "digital simplicity" with the expectation that such a concept would lead to an uncomplicated digital repository, which could be implemented and maintained, by small group of talented individuals with a moderate level of technical expertise in a variety of situations. This paper re-examines the notion of digital simplicity and considers whether such an idea is achievable or whether it is nothing more than a form of technical nostalgia, seeking for an imagined simpler times, or the creation of a digital ghetto, with sub standard functionality. It also considers what remoteness means in the digital world, and what are the implications for the systems that exist in such physical, geographical and technical environments. This paper will draw out issues that arise in the seemingly contradictory situation where interaction with the complex is a necessity of participation, but simplicity becomes a requirement of sustainable digital preservation.

Author

Kevin Bradley has worked in the field of oral history, sound archiving and digital preservation for just over 25 years, primarily for the National Library of Australia (NLA). Currently Curator of Oral History and Folklore and Director of Sound Preservation, Kevin's previous appointments include Sustainability Advisor on the Australian Partnerships for Sustainable Repositories (APSR), a DEST-funded partnership, Manager of the Sound Preservation and Technical Services, and Acting Director of Preservation at the NLA. Kevin has been extensively involved in the preservation of general digital objects. For a number of years he managed Digital Preservation at the NLA and worked on the infrastructure project Australian Partnership for Sustainable Repositories. Kevin is a member of the UNESCO Memory of the World Programme, Sub-Committee on Technology.

1. Introduction

In 2007 a paper was published on behalf of UNESCO's Memory of the World Subcommittee on Technology. The paper was entitled *Towards an Open Source Repository and Preservation System*, and it surveyed the open source repository software area, at least as it was known to the author and the research team at the time, to determine whether a narrowly focused, fully OAIS functional low-cost repository system could be developed. The need for this is very clear; all over the world the preservation, digitisation and documentation of social, cultural and technical knowledge is being undertaken in the

digital domain. It is widely accepted that digital technology in isolation is a fragile media, and proper systems and technologies are required to maintain the content with integrity in both the short, medium and long-term. It is also equally obvious that the technology and systems to maintain that data in the medium and long-term exists primarily in more wealthy countries with large technical infrastructures. The paper stated that such an aim was possible.

This paper, like its predecessor, argues that we have a responsibility to develop appropriately scaled, responsibly safe, digital repository systems with preservation capability in the long-term, and data security in the short and medium terms as being one of its prime criteria. It also suggests that the technology and information wealthy countries should develop open source, freely available, supported solutions for the issues described in order to enable nations, communities and others to exercise the right and ability to choose whether they wish to maintain and manage their own heritage materials in the digital domain.

2. The Urgency in the Need for a Repository.

All digital information, if it is to be managed and maintained, must be stored in an appropriate system, however there are a number of categories of digital materials with different urgencies associated with the need for a repository. As has been widely discussed, the prime motivation for the digitisation of most two-dimensional physical media is access. And as most digitisation in the current digital world is a form of access driven image digitisation, most of the tools required for dealing with digital information needs are about organisation for access. Libraries and archives throughout the world have been creating digital images of their paper-based collections in order to make the material more freely and widely available, or at least to facilitate local access without damage to original materials. This process has a number of benefits, not least of which is a wide ranging and democratic access to information. The repositories and other digital tools that manage this are designed to simplify the processes and maximise the availability and discoverability of content.

Consequently, with regards to this particular scenario, the primary driver for building long-term data management capability into a repository is the economic benefit that might be enjoyed compared to the cost of read digitising an image collection. Even though it is also possible that in the long-term these image surrogates may well be the copies that out-survive their original sources, the risk of losing an image copy of a paper-based item and not being able to replace it in the medium term is quite low for most paper based material. And as long as paper-based records continue to survive in a robust form, digital repository managers are able to continue with their commonly risky data management policies in which the only likely loss is the cost of recreating the digital images, rather than the content itself. The risk based approach has effectively, if not consciously, been the way most small scale digitisation systems operate. As a consequence, the open source digital repository environment has not developed this capability as thoroughly as they have access and dissemination.

However, most sound and audiovisual documents in the 20th century were created electronically, and these machine-readable electronic records are now either obsolete or physically decaying due to the chemical failure of the carriers. The urgent need to copy and preserve the content of these items has been well documented and well described elsewhere[1]. In addition, most sound and audiovisual and still images

[1] Schüller, Dietrich. "Preserving the Facts for the Future: Principles and Practices for the Transfer of Analog Audio Documents into the Digital Domain," Journal of the Audio Engineering Society (Vol. 49, Nos. 7/8), July/August 2001.

created since the end of the 20th century have been created in digital form and stored on impermanent short term carriers. Most contemporary collections are acquiring original digital photographs, sound recordings and video, without a clear pathway to manage them. So, where these documents are important, and it is necessary to maintain them, the repository that holds this material will be responsible for the only copies of these fragile virtual items. The future will not have access to most of the original carriers, but will have access to the content only through digitised facsimiles of managed data or files of the original object.

Large-scale data repositories and the National collections in the developed world manage the data in the same way as the banking environment that pioneered the approach; that is sophisticated and expensive storage management systems in an integrated data environment using multiple types of media and backed up by an IT team or contractors to undertake those tasks. However, these large-scale systems do not tend to scale down, and are not common in the smaller scale, remote collections, or collections in underdeveloped or developing countries because of the cost and the local infrastructure. So the urgent need for an open source repository system that incorporates preservation quality data management and backup is because those repositories to maintain these vital documentary materials need to maintain them as the sole source of otherwise lost materials.

Also equally clear is that the repositories of digital information are only as trusted, or as trustworthy, as the institutions that maintain and house them. So while it is important to create the technology that will allow long-term preservation, this will only be efficacious where the institutions that own and create those documents take on that role for the continued maintenance and existence of the content of those materials. Even though this paper argues for a widely available, reliable and low-cost system that could be easily deployed in a remote or technically less sophisticated environment, it also recognises the limitations of that aim. For a repository to be sustainable it must exist in associated technical infrastructure that is capable of supporting a functioning and sustainable system. Similarly there must be some level of technical knowledge and lease some recurrent resources, albeit at lower level, to make it sustainable.

3. The Outcomes of the 2007 Report

The report, Towards an Open Source Repository and Preservation System, was based on a set of industry understandings, which by and large could be considered fairly common sense. Listed below are some of the points that informed the document. Underpinning these fairly broad statements was a plea for solving the digital preservation problem for small-scale repositories in a way that dealt with the majority of simple or basic digital objects. This statement was required because it was recognised that much of the work being done in digital preservation is solving the long-term sustainability issues for some very sophisticated technological objects. However the very complexity of the problems such expertise are trying to solve often excludes simpler systems and solutions which might be applied to the majority of

---

Bradley, Kevin. "Critical Choices, Critical Decisions: Sound Archiving and Changing Technology."Proceedings of the Pacific and Regional Archive for Digital Sources in Endangered Cultures, 2004.
Schüller, Dietrich ed. 'Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy Version 3", International Association of Sound and Audiovisual Archives. IASA Technical Committee
Bradley, Kevin, ed. "Guidelines on the Production and Preservation of Digital Audio Objects, IASA TC-04 Second Edition.": International Association of Sound and Audiovisual Archives-IASA Technical Committee, 2009

formats that are used to document and record cultural materials generally. In the case of the digitisation programme, the organisation has control over the formats and so even simpler solutions are potentially suitable.

- Create and store the content on a digital file in a format which does not apply any form of manipulation which causes data loss or loss of authenticity.
- Use a format which is widely implemented and supported, and preferably, though not necessarily, open or non-proprietary.
- Use a format that has a potentially long life (digitally speaking).
- Use a format that is most likely to have available migration pathways to the next format.
- Store enough metadata to be able to facilitate identification, access and preservation processes.
- Use a reliable storage format on at least two types of carrier.
- Make multiple copies, and check and verify them regularly.
- Plan to replace carriers and software as the market demands, and plan to migrate the content to the next type of reliable carrier.

The notion of digital simplicity, as discussed in the 2007 UNESCO paper, is appealing because it implies a simple long-term preservation plan. However, such simplicity would be too great a cost if it were at the expense of providing access to significant content. Rather I wish to show that the notion of digital simplicity is embodied in a series of concepts regarding formats, content and their relationships, which when applied to long-term digital preservation increases the likelihood of survival of the content embedded in these formats. Digital simplicity can be considered at two levels; the level of the object that is being managed; the level of the repository that is being deployed to manage that object.

One of the main complexities, and one that has frustrated many thinkers about digital information, is identifying the actual characteristics of the digital object that need to be preserved. When looking at complex, computer-based or web deployed digital objects, it is difficult to distinguish between the information the object carries, the media or format that carries it, and the manner in which that object presents the information it carries. The media is, after all, the message, and as a consequence detaching the significant properties, or the essence, of such complicated digital objects is difficult. When the meaning embodied in an object is embedded in the objects, the relationships between other objects, and the way those objects are presented, the complexity multiplies. Virtually all modern web publication fall into this loosely defined category of complex materials. The technical implications of this is that being unable to define precisely what is the essence of an object, it becomes nearly impossible to design the systems which manage the long-term preservation of those characteristics.

The National Archives of Australia describes a number of categories of material that it takes into its digital repository[2]. They are listed below. Note however that they describe categories of material from the point of view of format or encoding information, rather than user categories.

Archive files/wrapper/container

[2] Cunliffe, Allan "Dissecting the Digital Preservation Software Platform Version 1.0 RKS: 2009/4026" 2011 http://www.naa.gov.au/Images/Digital-Preservation-Software-Platform-v1_tcm16-47139.pdf Accessed August 2012

Audio
Computer aided design (CAD)
Email
Geospatial data
Image
Image - Vector
Office documents
Plain text
Database tables, such as comma and tab-separated files (csv, tsv)
Scripting files (such as Python, Javascript, Perl, PHP)
Structured Query Language (SQL)
Video – including video stream, audio stream and container

Of these, audio, image and video might be construed as simple digital objects, and text may in certain circumstances be the bridge between simple and complex objects. There are a number of ways of looking at this, of which the first may be with regard to significant properties. JISC defines Significant Properties as follows:

> Significant properties, also referred to as "significant characteristics" or "essence", are essential attributes of a digital object which affect its appearance, behaviour, quality and usability. They can be grouped into categories such as content, context (metadata), appearance (e.g. layout, colour), behaviour (e.g. interaction, functionality) and structure (e.g. pagination, sections). Significant properties must be preserved over time for the digital object to remain accessible and meaningful.[3]

The significant properties of an audio file is in fact its audio characteristics, that is, if the process sets out to preserve an audio file it must preserve the complete technical/quality characteristics that embodies it. Similarly, this could be said of still images, and to some extent moving images as well. The significant properties are technically embodied in the quantitative measure of their quality. Text based documents are complex symbolic items that human society has been creating and adding to over the millennia. While it is quite possible to show examples that embody that complexity, our long familiarity with such objects allows us to make some pragmatic decisions about font, layout and spacing, and so read and preserve some fairly basic digital objects in text form. However, there is recognition that for most text objects, its significant property would be the information embodied in the text rather than other characteristics.

All other items in the list are a complex mixture of text, technology, relationships, files, standards and components that seems to defy a simple and automated way of being managed. So a simple digital object may well be one that the significant properties can easily, and technically, be separated from the media that carries it.

In addition to this, most of the formats associated with the simple objects described above have had a relatively long life, in digital terms. So, for example, image and audio formats used for preservation have been stable and standard for the past decade and a half. This is because the professional industry tends to resist change, unlike the consumer market, which tends to embrace it. Likewise, professional involvement in the format means that there will almost definitely be a migration path from the old format

[3] "The significant properties of digital objects" last modified 17 August 2010 http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops

of the new professional format; otherwise the market providers would find it difficult get the buy-in of the professional user.

These formats, all these categories of material, are often the primary source materials from which other more complex objects are created. They are also most often the primary documents for which most archives collections have the greatest responsibility, which carry the most significant information. So, a repository that exhibits digital simplicity would be one that deals with a limited number of files that exhibit characteristics described above.

Digital simplicity does not mean a technical ghetto, or a backwater that remains the same while the rest of the world changes, rather it implies an appropriate technology for the preservation of significant representational items which fall into a defined category, and which are the most significant for many archives.

The Towards an Open Source Repository and Preservation System (2007) report made a recommendation for UNESCO's involvement "in open source software development on behalf of the countries, communities, and cultural institutions, who would benefit from a simple, yet sustainable, digital archival and preservation system", and to this end the Information for All Programme (IFAP), at the recommendation of the SubCommittee on Technology (SCoT), provided funding for the development of certain functionality, for which Archivematica won the project.

The Open OAIS defines Data Management, Ingest, Access, Administration, Preservation Planning and Archival Storage. The report also noted that most repository software was well developed in the areas of data management, ingest, access and administration, but little no work had been done in preservation planning and archival storage. The intention of the 2007 report was to create a repository that included the potential to manage all six of the OAIS categories.

To this end of the work undertaken by Archivematica produced a set of repository tools that met those criteria in five of the six OAIS categories. Perhaps most significant was the inclusion of PREMIS metadata which enabled preservation planning in a way that was not possible with other open source systems. However, data storage remains an unresolved issue.

Most large-scale repositories buy large-scale data storage systems, and enter into some sort of arrangement with companies to supply the technology, software and data tapes. As has been previously stated, this sort of arrangement does not scale down well to a small-scale repository, represents a significant expense even on large, comparatively well resourced repositories and is much more expensive per gigabyte of storage when the repository data size is not great. The 2007 report looked to find a cost effective alternative.

At the time of writing (2007) the report stated that AMANDA, (Advanced Maryland Automatic Network Disk Archiver), an open source system which allows the IT administrator to set up a single master backup server to back up multiple hosts over network to tape drives/changers or disks or optical media, had potential to be incorporated for this purpose. Since that time the AMANDA system has been developed significantly and so has even more potential if it were incorporated into a repository system.

However, the data management landscape has also altered significantly, and it is worth reviewing at this time the options for managing data in a small-scale, tightly resourced archival digital repository.

There are a number of technical and socio-technical ways that a tightly resourced repository can ensure the medium to long-term integrity of the data it holds. We will consider some of those below; repository management partnerships, third-party providers, cloud storage, disk only and disk/tape systems. However, it is important to recognise that no media will last forever, and no data storage technology is reliable, rather it is the technical system that incorporates the technology that provides the

reliability. The second aspect to recognise is that all data storage systems are at heart, a risk management technology, and the task is to find an appropriate technology which addresses appropriately the level of risk for the content the system manages and preserves and the environment in which it operates. All these technical systems are based in a particular environment, and we have to interact with that broadly defined system to fulfil all the tasks of a digital repository: collection building and management, preservation, and dissemination. In order to retain and sustain our collections we must create a continuous link between the present and the future, a link that ensures that the ongoing links in the chain of sustainability create a digital object that the future user can access in just the way we do now and does so in the specific social, political and economic circumstance in which it exists.

The archives, libraries and other knowledge keepers are dependent on these sophisticated technical systems and networks that are the pinnacle of our socio-technical capabilities. This means that the future of our digital and digitised materials are linked directly to the ongoing and continuing support of the government and other organisations which fund our archives and collections. It is no overstatement to say that we are linked to a continuing civilisation.

More importantly, or perhaps just more urgently, we are linked to the ongoing economic and social stability of the societies in which we live and at risk when that stability is threatened. In the current, and seemingly ongoing international financial crisis, it is very concerning to realise that the greatest risk to our collections is not technical or chemical, is not related to magnetic domains, to optical surfaces or chemical solutions, but rather to the economic viability of the societies in which we live. Without ongoing funding, infrastructure and support, our collections are at risk. When considering the appropriate approach to collection management and preservation, we must take note of the social, cultural and economic environment in which that technical system operates. We must create systems that can be sustained in the economies in which they operate.

Repository management partnerships: it may well be appropriate to take all data held in a repository and back up that content in a partner repository that has the highly reliable, low risk, well resourced solution to the management of content. This effectively moves the risk out of the local storage environment and into the storage environment of another, remote place. The issues associated with this are by and large not technical, but rather those of the ownership of content and the property at cultural rights associated with it. Many national or nationalistic institutions tend to resist this, feeling quite rightly that they should be able to manage their own cultural information. Nonetheless, a very good relationship also has the effect of storing data in two quite separate areas, probably separate countries, which has an enormous advantage in protecting content against disasters. However, once crossing national borders and outside of national jurisdiction, other laws may apply to the content with regard to rights and ownership. Trust is probably the most significant aspect of this sort of relationship. However, when it comes to economic motivation, one of the flaws in the approach is that the repository that has the responsibility to maintain and manage the material is not the one that has the motivation to do so.

Third-party providers: There are an increasing number of commercial providers who will manage storage of digital content. Some of these provide very high quality, reputable systems, that replicate those found in the major national archives and libraries around the world, or indeed, exceed their capabilities. Some do not have those types of technical redundancies and capabilities. Besides assessing whether such suppliers meet the technical requirements of the repository, and whether the content they provide is managed with appropriate regard for the ownership of the original content and the legal constraints of the

country from which they came, there is a need to negotiate a contract shall agreement regarding the quality of the data, an export and retrieval requirement in the case where either the storage company fails, or the content owner is unable to pay storage fees.

Technical solutions: "May all your problems be technical" Ancient I.T. worker blessing.

Cloud storage is the distribution of data and information across a number of storage environments, which are accessed, managed and controlled through a wide area network. The appearance of a single storage environment is created through a virtualiser, which keeps track of all the parts of the storage environment and renders it as a virtual storage system. Typically there is, or should be, a very high level of redundancy to manage the complexities of multiple storage environments. As can be seen, Cloud storage does not use any new storage technologies, but rather exploits the availability of a widely distributed high bandwidth network and uses the same storage technology as everyone else. Cloud storage hides the fixed infrastructure and allows the marketplace to respond more flexibly to changes in demand for data storage. Cloud storage also can enable total cost savings because the fixed costs of maintaining a system are spread across multiple users.

However, the security risks associated with data stored in multiple environments, and discussed above for partnership storage, is multiplied by the number of storage environments, the crossing of national boundaries, and the transmission of that data across multiple environments. Like all aspects of data management these risks need to be considered and managed.

The same architectures and data models that underpin the OAIS apply to data stored in the Cloud in the management of long-term preservation of digital objects. There are also management risks and The National Archives of Australia have released a document entitled "A Checklist for Records Management and the Cloud"[4] which highlights issues around security, authenticity, completeness, discoverability and access, and variability due to data management, copying and migration. David Rosenthal's blog about the announcement of Glacier, and Amazon based storage environment, raises issues about latency and the cost of access over a given access threshold of around 5% per month (which means that retrieving an entire archive may take a long time, and/or cost a significant amount of money).

However, the most significant issue for an archive situated in an environment with limited infrastructure is the bandwidth of the connection to the network. The objects stored in digital repositories for long-term preservation tend to be very large, especially for audio and video items, and these items require high-speed large bandwidth networks to distribute the content around. As has been discussed widely, the connection speeds in developing countries can be very slow, and may be unreliable. In these circumstances a Cloud storage environment will be totally impractical.

Disk only storage: there is an increasing opinion that storing data on disk (spinning disk/hard metal) in RAID arrays is suitable for collections of significant and important data. Certainly hard disks manufactured in the past few years have better specification and performance characteristics than those from an earlier period. Nonetheless, numerous papers point to the failure of disk storage in critical environments, and the likelihood of data loss.[5] Complex modelling of real-world failures point to a likely

[4] A Checklist for Records Management and the Cloud, National Archives of Australia, 2011. Last modified 2012 http://www.naa.gov.au/records-management/publications/cloud-checklist.aspx

[5] Harris, Robin "SSDs vs. Disks: Which Are More Reliable?" Last Modified January 2011 http://www.datacenterknowledge.com/archives/2011/01/27/ssds-vs-disks-which-are-more-reliable/

loss of data even when stored on disks configured in RAID. In big arrays, the likelihood of data loss is even greater as the multiple failure and replacement occurrences increased the likelihood of two failures in a single array. The conclusion of many such forums and discussions is to store your data in at least three different locations and to continue with the old advice of different types of media. It is also salutary to observe that most major archives still use tape and disk to maintain the integrity of the data.

Disk and tape, with multiple copies on tape, meets the requirements described in "Towards an Open Source Repository and Preservation System", and remains a requirement for the small scale, tightly resourced repository described therein. As a consequence, there is a need to find funds to develop the incorporation of this into standard open source repository systems. The requirement is specific to those repositories in developing countries for all the reasons described above. However, it is not a requirement that is likely to be implemented for a well resourced repository even if it is using the same open source repository software. This is because well resourced repositories in the economically leading countries tend to be implemented and deployed within existing technical systems, using existing IT infrastructure and staff expertise. As a consequence, developing country is implementing open source repository software will not be able to piggyback off the developments required by their richer counterparts. As most repository deployment in these environments also includes new technical infrastructures, the cost barrier to establish a reliable sustainable repository is more significant in these cases.

4. Conclusion

There is still a need, especially in small scale, tightly funded digital repositories in archives in developing countries to find a solution to medium term data integrity, and to build that into the open source systems that are available. However, because the primary motivation for this comes from an underfunded sector, there will be a need to find grant funding to implement this necessary requirement as it will not come from the commercial sector as a response to its need, not the market generally.

Harris, Robin "Google's Disk Failure Experience" last modified February, 2007
http://storagemojo.com/2007/02/19/googles-disk-failure-experience/
Harris, Robin "Everything You Know About Disks Is Wrong" last modified February, 2007
http://storagemojo.com/2007/02/20/everything-you-know-about-disks-is-wrong/
Pinheiro, Eduardo, Wolf-Dietrich Weber and Luiz Andre Barroso "Failure Trends in a Large Disk Drive Population" Google Inc in the Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07), February 2007
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/disk_failures.pdf